RESEARCH BRIEF \ JANUARY 2018

Results from the APT Validation Study III:
Reducing cultural biases in youth program observations

By Amanda Richer, M.A., Ineke Ceder, and Linda Charmaraman, Ph.D.

Background

The first Afterschool Program Practices Tool (APT) Validation study provided evidence that the APT has many strong technical properties, such as short-term stability and a strong factor structure, and is an appropriate tool for measuring afterschool program quality. However, this work also showed that rater reliability was somewhat unstable.

The second APT Validation study introduced the use of video segments to establish master-coded ratings to help calibrate raters and improve rating accuracy. Results from the second study showed low average accuracy scores, ranging between 51 percent and 58 percent, and revealed that Black participants had consistently lower accuracy scores compared to White participants.

Wellesley Centers for Women

Current Study

The primary goal of APT Validation Study III (2016-2017, funded by the William T. Grant Foundation) was to examine the reasons behind the Black/White scoring gap and to eliminate cultural biases in assessment. To eliminate the scoring accuracy discrepancy we reviewed the selection and narrative framing of the video clips, and included master scores to the final scoring guide.

Research objectives

1. To develop APT reliability exams in which the average rater score falls within the field benchmark of 80 percent;
2. To develop and refine APT reliability exams without potential for cultural bias, and examine whether other demographic factors are associated with better performance on reliability exams, such as gender, region of the country, number of years of OST experience, experience with external program evaluations;
3. To determine whether: (a) familiarity with the Anchors Guide, (b) experience using APT generally and as an external evaluator, and (c) APT training are positively associated with better performance on reliability exams.

Master Scoring Procedures

A diverse group of female APT experts (three African Americans and one Latina) was recruited to provide master scores for 35 selected video clips. They received guidance on how to reduce cultural biases during this process. After each of them entered scores rating all clips, a fifth consultant (White male) was assigned as a "tie breaker" for those clips which did not receive consensus from at least three out of the four initial coders. If the fifth consultant did not tip the number of identical ratings over 50 percent, the clip was discussed at a meeting to reach full consensus or to be dropped. Reasons articulated for the master scores during initial coding and at the consensus meeting were recorded and used to develop practice exams for this study.

Exam Participant Description

Our participant group took three exams and three practice exams, which they were assigned in random order. The group consisted of 48 raters from 11 states and was 33 percent non-White, which represents the most geographic and racial/ethnic diversity among participants in all the previous APT validation studies.

The majority was female (77 percent) and White (67 percent) in addition to 17 percent Black, 10 percent Hispanic, 4 percent Asian, and 2 percent Native American. Nineteen percent of the sample were between the ages of 20-29, 38 percent between 30 and 39, and 44 percent were 40 and older. Most were employed in afterschool organizations in the Northeast (73 percent) or the South (21 percent). Most were familiar with the APT Anchors (73 percent) and almost half report using the APT one to two times per year.

As for APT training, 79 percent reported receiving in-person NIOST-based training (National Institute on Out-of-School Time), 52 percent reported receiving online NIOST-based training, 25 percent reported being trained at their own site, and 27 percent were trained through a previous APT Validation Study. Raters could report more than one type of training.

Data Analysis

Item-level

Overall, most participants rated an exact match to the master score across items. However, a few items per exam had very poor accuracy scores and were removed to ensure that the exams would be practical for the field. In order to mitigate that one best "accurate" quantitative score is assigned to an observational rating that is rather qualitative in nature and subject to personal biases and also in an effort to reduce chances of one point of view becoming the "gold standard," we decided to allow two accurate ratings for the majority of items, such that a "true" rating can land in between two scores. In fact, this was an observation frequently made during master scorer discussion. These decisions fall in line with other observation scales in the field which do not measure reliability with exact agreement with master raters, such as for example Teachstone's CLASS instrument wherein an accurate rating falls within one1 point on either side of the master code (i.e., admitting three accurate ratings; see Bell et al., 2012).

Rater-level

A percentage correct score was calculated for each participant for each exam to assess accuracy levels. One-way analysis of variance and independent samples t-tests were used to assess significant group differences in rater accuracy.

Results

Research Objective 1.

The average rater accuracy score was 82.4 percent for Exam 1, 84.9 percent for Exam 2, and 86.5 percent for Exam 3 reaching our goal of 80 percent for all exams.

Research Objective 2.

Group difference tests of rater accuracy by gender, race, age, region, education background, and out-of-school time (OST) experience were conducted on the total accuracy scores and 80 percent passing benchmarks for each exam. No significant differences were found when comparing accuracy scores of Whites and non-Whites, males and females, across age groups, those residing in New England versus those who reside outside New England, and those with K-5 OST experience versus those with no K-5 OST experience. Results showed no significant

Results, continued

group difference in accuracy for participants who had experience working with minority students, low income students, students in urban environments, and staff working in large programs with high student-to-staff ratios compared to raters who did not report having these experiences, suggesting that the exams demonstrate no bias toward whether raters have or have not worked with vulnerable OST populations. For educational background, we found an overall significant group difference for total score on Exam 2; post-hoc analyses showed that participants with a PhD were less accurate compared to those with a four-year bachelor's or master's degree.

Overall, we found no significant group-level biases in passing rates across the three exams, indicating that the reliability exams were designed to not favor one type of rater over another.

Research Objective 3.

Familiarity with APT anchors. We compared whether familiarity with the APT anchors (responses=yes or no) influenced accuracy scores, and found significant differences for all three exams. For Exam 1, raters who were not familiar (N=13) with the APT anchors were less likely to pass the exam at the 80 percent benchmark compared to those who were familiar (N=35) with the APT anchors. For Exams 2 and 3, raters unfamiliar with the APT anchors had lower total accuracy scores and also were less likely to pass the exam at the 80 percent benchmark. For all results, having familiarity with the Anchors was associated with better accuracy.

Frequency of APT usage. Raters reported how often they used the APT per year which ranged from never to five or more times per year. We compared this grouping across accuracy levels for each exam and the 80 percent passing benchmark, and found no significant differences across groups for Exams 2 and 3. However, for Exam 1 we found a significant group difference between raters who used the APT three to five times per year and those who used it five or more times per year. Those using the APT more often were more accurate on Exam 1; however, passing Exam 1 at the 80 percent benchmark was not significantly different across the groups.

Using APT for evaluation. We found no significant differences in accuracy for those who had used the APT solely in their program versus those who did not for Exams 1 and 3. However, we did find that participants using the APT outside of their own program as part of an external evaluation (94 percent with external experience vs 72 percent with no external experience) were more likely to pass Exam 2 at the 80 percent benchmark.

APT Training. Participants reported what type of APT training they had received throughout their experience and how long ago they received this type of training. Participants could report having received in-person training, online training, afterschool program administered training, and training conducted through previous APT research studies. We explored potential training effects in a variety of ways. We compared participants who had completed a particular type of training within the past year versus those who did not, we compared participants completing a training type versus those who never completed the type of training, and lastly we compared groups by the number of trainings they completed. We did not find any significant group differences in accuracy when comparing the recency of training and whether participants had or had not completed a type of training. We did find a small overall significant difference for Exam 1 for passing at the 80 percent benchmark and the number of trainings completed. Significant differences were between participants who completed two trainings (89 percent passed) and those completing three trainings (38 percent passed). Results for the remaining groups were 61 percent passed with one type of training and 100 percent passed with all four types of training. Incidentally, of those who had completed two trainings, a majority completed online and in-person NIOST training.

Conclusions

The results from this study show that in all three reliability exams raters were able to achieve accuracy passing levels of 80 percent. We did not find any significant differences in rater accuracy across the three exams, suggesting the exams are equivalent in assessing rater accuracy using the APT Activity sections. We found no significant passing rate differences by rater characteristics, such as race, gender, age, region, and experience with OST populations. Lastly, there is some evidence to suggest that familiarity with the APT anchors, higher frequency in using APT, and using APT for external evaluation purposes is related to higher accuracy scores. Since knowledge and frequent use of APT anchors are intermingled with having been previously trained to use the APT, the exact relationship between rater accuracy and the need for a particular type of training still needs further evaluation. We recommend further development of a reliability training particularly for higher stakes raters, for example in the form of an advanced training tool for current APT users to gain expert proficiency, knowledge, and practice using the APT anchors to guide program quality ratings in multiple external program settings.

Author bios

Amanda M Richer, M.A., is a research associate and associate methodologist for the National Institute on Out-of-School Time and the Wellesley Centers for Women. She has been involved in the psychometric testing of the APT tool since the first APT Validation Study and continues to invest her time in using data to promote and enhance OST tools.

Ineke Ceder is a research associate at the Wellesley Centers for Women. Most of her work has focused on gender and race, women's leadership, and sex education. Her involvement on all three validation studies of the APT Tool has driven home for her, once again, that research data drive social change and can bring equity for all.

Linda Charmaraman, Ph.D., is a research scientist at the Wellesley Centers for Women. As the co-PI of APT II and PI of APT III, she has solidified her strong engagement in how research evidence can inform policy and practice as well as how researchers can reduce structural inequality gaps in all areas of their work.

Acknowledgements

This research was generously funded by an Officer's Research Grant #187010 from the William T. Grant Foundation. Publication support was also provided Dr. Charmaraman by New Connections: Increasing diversity of RWJF programming at the Robert Wood Johnson Foundation. We would like to thank Lisette DeSouza, Ph.D., for her contributions to this project and to our study participants who made this research possible.

References

Bell, C., Gitomer, D., McCaffrey, D., Hamre, B., Pianta, R., & Qi, Y. (2012). An argument approach to observation protocol validity. Educational Assessment, 17, 62-87.

Kishida, Y., & Kemp, C. (2010). Training staff to measure the engagement of children with disabilities in inclusive childcare centers. International Journal of Disability, Development and Education, 57(1), 21-41.

Schlientz, M.D., Riley-Tillman, T.C., Briesch, A.M., Walcott, C.M., & Chafouleas, S.M. (2009). The impact of training on the accuracy of Direct Behavior Ratings (DBR). School Psychology Quarterly, 24(2), 73–83.